PR-04-01

ANALYSIS OF THE STALLING AND FLOOD INCIDENTS ON THE FEBRUARY 2004 ADMINISTATION OF THE CALIFORNIA BAR EXAMINATION

Stephen P. Klein, Ph.D. and Roger Bolus, Ph.D.
GANSK & Associates

May 17, 2004

Summary

Two incidents occurred on the last day (Thursday) of the February 2004 administration of the California bar exam. One incident involved the roughly 35 percent of the applicants who answered essay questions 4–6 on their laptop computers. Every four minutes these applicants' computers automatically saved the keystrokes they had made during the previous four minutes. However, the keystrokes they made while the auto-save process was underway did not appear on their computer screens until the auto-save was completed. This "stalling" only occurred on Thursday morning. It did not occur on Tuesday morning when the applicants answered essay questions 1–3 or on the Performance Test (PT) sections of the exam. Laptops are not used on the multiple-choice section.

The time it took to complete an auto-save and thereby the length of the stalls increased as the test session progressed. Stall time also was a function of the applicants' editing styles and their computers' hardware characteristics. Consequently, some applicants accumulated much more stall time than others.

The laptop users' mean score on an essay question was about one point higher than the non-laptop users' mean. This was true on Tuesday when there was no stalling and on Thursday when there was stalling. In addition, both groups had a three-point higher average essay score on Tuesday than on Thursday. Thus, after controlling for the difference in average question difficulty between days, the laptop users did just as well when the stalls occurred as when they did not occur. It therefore appears that the stalls did not affect the laptop users' scores.

Further analyses revealed a slight (and statistically insignificant) positive relationship between Thursday morning scores and how much stall time a candidate encountered; i.e., on the average, the laptop users' scores *increased* slightly as their stall times increased. The only exception to this trend was with candidates who averaged more than five minutes of stall time per hour. Their Thursday morning scores usually decreased slightly as their average hourly stall time increased.

A regression analysis (that considered the applicant's stall time and scores on the rest of the exam) indicated that the appropriate adjustment for this decrease would be to add 0.01 points to the applicant's score on question 6 for each second of average hourly stall time the applicant experienced in excess of 300 seconds (five minutes). For example, an applicant who experienced an average of 400 seconds of stall time per hour of testing time would receive a total of (400-300) x 0.01 = 1 additional point on the essay section.

The applicants who switched from using their laptops to hand writing during the Thursday morning essay session also had about a 3-point higher mean on Tuesday than on Thursday. Thus, they did just as well on essays 4–6 as would be expected given their essay 1-3 scores and the pattern with other applicants.

The other incident with the administration of the February 2004 exam was a flood at the Pasadena laptop test center that delayed the start of its Thursday morning session. A regression analysis found that this delay did not adversely affect scores at this site. However, it did force canceling the administration of the Performance Test B (PT-B) task that was scheduled for Thursday afternoon.

Two methods (regression and pro-rata) were examined for imputing PT-B scores for the Pasadena applicants who had this session cancelled. This analysis found that these methods yielded identical passing rates. They also agreed almost perfectly in which candidates they would pass. Given these findings, we recommend using the pro-rata method because (1) it can be used with those taking the Attorneys' Examination as well as those taking the General Bar Exam and (2) it is consistent with the procedures the Committee used several years ago to estimate missing scores when an earthquake led to canceling the afternoon portion of the multiple-choice section of the exam at one site.

The pro-rata method consists of multiplying an applicant's reader-assigned PT-A score by 2, adding this product to the sum of an applicant's scores across essay questions 1 through 6, and then dividing the sum of these scores by 7.86. The 7.86 scaling factor in this equation adjusts for differences in mean raw scores among the different parts of the written (essay + PT) portion of the exam.

Overview

The next portion of this report reviews some important features of the California exam and the rules for computing scores. We then discuss the stalling that occurred on Thursday morning, the analyses that were conducted to investigate the impact of these stalls, and the results of these analyses. We also describe how the scores of those with relatively long stall times might be adjusted. Next, we report our analyses and findings regarding the flood incident. The last section summarizes our conclusions and recommendations.

There are three appendixes. Appendix A lists the members of the panel of experts who provided advice regarding the methods we used to investigate the computer software and flood incidents described above. Additional information about these methods is presented in Appendix B and C.

Exam Structure

California's three-day General Bar Exam consists of six essay questions, two Performance Test (PT) problems, and a 200-question multiple-choice test called the Multistate Bar Examination or MBE. Table 1 shows the test administration schedule.

Table 1
February 2004 Test Administration Schedule

Test Session	Tuesday	Wednesday	Thursday
Morning	Essays 1–3	MBE items 1 – 100	Essays 4–6
Afternoon	PT-A	MBE items 101-200	PT-B

Each test session is three hours in length. However, several applicants received testing accommodations that included allotting them extra time. Essay questions are not timed separately and applicants can answer the three questions within an essay test session in any order they choose. They also can go back and forth between questions within a test session.

On the essay and PT portions of the exam, applicants have the option of hand writing their answers, preparing them on a typewriter, responding on their laptop computers, or using some combination of these response modes. Applicants who prepare their answers on laptops must (for test security and fairness reasons) use a software program that prevents them from accessing their hard drives during the test session.

The attorneys who grade the essay and PT answers assign scores in 5-point increments on a 40 to 100-point scale. The reader-assigned PT scores are then multiplied by two. The maximum possible written raw score is therefore 1000 points (six essays at 100 points each plus two PT problems at 200 points each = 1000 points).

Total written raw scores are converted to a score distribution that has the same mean and standard deviation as the applicants' MBE scale scores. Total scale scores are computed using the formula below. Applicants need a total scale score of 1440 or higher to pass.[1] However, attorneys who have practiced in another jurisdiction for at least four years can opt out of taking the MBE. The lawyers who select this option (which is called the "Attorneys' Examination") can pass the bar exam by earning a written scale score of 1440 or higher.

Total Scale = (.65 x Written scale) + (.35 x MBE scale)

Computer Software Incident

The laptop users utilize a software program that automatically stores a copy of their exam file every four minutes throughout a test session. This program was modified for the February 2004 exam. An unanticipated consequence of this change was that the keystrokes the applicants entered while an auto-save was underway did not appear on their screens until the auto-save was completed. This "stalling" only occurred on Thursday morning when they were answering essay questions 4–6. It did not occur on Tuesday or on Thursday afternoon.

The length of a stall was a function of how long it took a candidate's computer to copy and store that candidate's exam file. The time it took to do this became longer as the test session progressed (i.e., as the total number of keystrokes to be saved accumulated). Stall time also was related to the characteristics of the applicant's computer hardware (laptops with faster processors and more memory had less stall time) and editing style (holding down the delete and backstroke keys to erase large blocks of text increased the length of the stalls).

Analysis Sample. Except where noted otherwise, the analyses below were conducted with all of the 1486 laptop users and 2877 non-laptop users (hereinafter referred to as "laptoppers" and "non-laptoppers," respectively) who took the MBE and had scores on PT-A and essay questions 1–6.

Response Mode Analyses. On the average, the difference in mean scores between the laptoppers and non-laptoppers on a Tuesday morning essay question (i.e., when the stalls did *not* occur) was the same as the difference between their means on a Thursday morning question (i.e., when there was stalling). In short, the difference in mean scores between the laptoppers and

[1] Applicants who come close to passing after the first reading of their answers have those answers graded again, but by a different reader. The results of the second reading are averaged with those from the first reading. If the applicant's total scale score after the second reading is between 1412 and 1439, then that applicant's answers are reviewed by a member of the Committee's Board of Reappraisers who makes a final pass/fail decision. All the analyses in this report are based on scores from the first reading.

non-laptoppers was not related to whether or not stalling occurred. Thus, the stalling did not appear to have an overall effect on scores.

Table 2 documents these findings. It shows that the average score on a Tuesday morning essay question was 64.7 for laptoppers and 63.6 for non-laptoppers; i.e., a difference of 1.1 points. The laptopper and non-laptopper means on a Thursday morning question (61.7 and 60.6, respectively) also differed by 1.1 points. Thus, on both days, the laptoppers scored slightly higher than the non-laptoppers. Both groups also scored an average of three points higher on Tuesday than on Thursday, but this difference was not related to an applicant's response mode. Hence, it must have been due to differences in average essay question difficulty and/or reader standards between days.

Table 2
Mean Score on an Essay Question by Test Session and Group

Questions	Stalling?	Laptoppers	Non-Laptoppers	Difference
Essays 1–3	No	64.7	63.6	1.1
Essays 4–6	Yes	61.7	60.6	1.1
Difference		3.0	3.0	**0.0**

Note: Standard deviations ranged from 5 to 7 points. Means were computed *before* any adjustments were made to any laptopper's essay scores.

To further investigate any overall effect, we constructed a regression equation to predict an applicant's score on essays 4–6 on the basis of that applicant's MBE, essay 1–3, and PT-A scores. This equation also had a term for whether the applicant used a laptop (coded "1" if the applicant used a laptop and "0" if the applicant did not use it). This analysis found that a laptopper earned about one-tenth of a point less per question on Thursday morning than would be expected. This tenth of a point difference was not statistically significantly different than zero (see Model 1 in Appendix B). In short, a laptopper's score on essay questions 4–6 was not significantly different than a similarly situated non-laptopper's score on these questions (where "similarly situated" was defined in terms of the applicants having comparable MBE, essay 1–3, and PT-A scores).

Applicants who switched response modes during the Thursday morning session did just as well on questions 4–6 as would be expected on the basis of their scores on questions 1–3. This finding is based on the following: (1) for both the laptoppers and non-laptoppers, their mean Tuesday morning score on an essay question was three points higher than their mean on a Thursday morning essay question and (2) there also was about a three-point difference in mean essay question scores between Tuesday and Thursday in the group of 68 applicants who used their laptops on Tuesday for questions 1–3 and then hand wrote their answers to one or two (but not all three) of the Thursday morning questions. The

Tuesday and Thursday means in this group of 68 applicants were 64.3 and 61.4, respectively; i.e., a difference of 2.9 points.

The results above indicate that on the average, an applicant's score on a Tuesday morning question was three points higher than that applicant's score on a Thursday morning question. This was true for laptoppers and non-laptoppers alike. Hence, the 3-point difference in mean question scores between test sessions had nothing to do with which response mode the applicant used or with whether the applicant changed response modes during the Thursday morning test session.

Measuring Stall Time. The stall time associated with a four-minute auto save is the number of seconds that elapsed while the candidate's computer processed the auto save. The candidate's computer screen was unable to display the keystrokes the candidate entered while this auto save was underway. However, the keystrokes the applicant made during an auto-save were recorded and stored as part of that candidate's exam file. These keystrokes appeared on screen at the conclusion of each auto save.

The analyses below examined whether the average amount of stall time an applicant experienced per hour of allotted testing time was related to the sum of that applicant's scores on essay questions 4 through 6. For these analyses, an applicant's *total stall time* is the sum of that applicant's stall times (in seconds) across all of the four-minute auto saves that occurred in the test session. The data required to calculate stall time was automatically recorded in the answer file on the diskettes the laptoppers turned in at the end of each test session.

Effective stall time was defined as the total stall time minus all the stall times associated with those automatic saves where there were no keystrokes in the four minutes preceding the save. In other words, the computer was on, but the candidate was reading a question, using the restroom, hand writing an answer, or otherwise not entering keystrokes (and thereby not experiencing stalls).

If a candidate struck any key at least once every four minutes throughout the test session, then that applicant's total stall time would equal the candidate's effective stall time. Total and effective stall times would be different only if the candidate stopped making keystrokes for at least four minutes during the test session. In short, total stall time corresponds to what the candidate's computer experienced whereas effective stall time reflects what the candidate experienced. There was a very high ($r = .94$) correlation between total and effective stall times.

Standardized stall time was computed by dividing the effective stall time (in seconds) by the total number of hours the candidate was given to answer the three Thursday morning essay questions. Three hours was allocated to all candidates except for those who received testing accommodations that included

more time (such as 4.5 hours). Thus, standardized stall time is the average effective stall time a candidate experienced per hour of allotted testing time.

Analyses Using Standardized Stall Time. If the stalls adversely affected a candidate's essay 4–6 scores, then it seems likely that the candidates who accumulated relatively long periods of stall time would have lower essay 4–6 scores than those who had relatively little or no stall times. We conducted several regression analyses to examine if this occurred. All of these analyses controlled on the applicant's MBE score, PT-A score, and total score on essays 1–3.

These analyses found that essay 4–6 scores did not decrease as stall times increased; i.e., applicants with above average amounts of stall time did not have lower essay 4–6 scores than similarly situated candidates with below average stall times (where "similarly situated" is defined as having comparable MBE, PT-A, and essay 1–3 scores). Indeed, essay 4–6 scores actually increased very slightly (and statistically insignificantly) as stall time increased (see Model 2 in Appendix B). In short, an increase in stall time was generally *not* associated with a decrease in essay 4–6 scores.

The foregoing results held for the laptoppers who had less than 300 seconds (5 minutes) of standardized stall time. For the remaining 12 percent of the laptoppers, there was a modest but statistically insignificant trend for longer stall times to be associated with lower essay 4–6 scores. While there is no compelling evidence of essay 4–6 scores being depressed by stall time, Regression Model 3 in Appendix B suggested that a reasonable adjustment would be to add 0.01 points to each candidate's score on question 6 for each additional second in standardized stall time over 300 seconds that the applicant experienced.

For example, suppose an applicant had 500 seconds of standardized stall time. This is 200 seconds more than 300. According to Model 3, this candidate should receive 0.01 x 200 = 2 additional raw score points. Because of operational requirements, all adjustments have to be made in whole numbers. We therefore recommend rounding fractional values to the nearest integer and adding the adjustment to the raw score on question 6.

When we used the applicants' February 2004 MBE and first read essay and PT (or imputed PT) scores to estimate the effects of this proposed adjustment, we found that one more applicant in our analysis sample would pass. One reason there was not a greater impact was that many of the applicants with relatively long standardized stall times passed even before the adjustment was applied (including those with over 10 minutes of standardized stall time).

Sensitivity Tests. The findings above did not change when we used *total stall time* instead of *standardized stall time* in the regression equations. For example, when we replaced standardized stall time with total stall time in Model 2 in

Appendix B, the coefficient for total stall time was 0.000455 (which is positive and not statistically significant).

We also examined the sensitivity of the adjustment suggested by Model 3 by investigating what would happen if candidates were awarded 0.02 points for each second over 300 seconds of standardized stall time; i.e., double what the regression analysis suggested. This produced exactly the same passing rate as the 0.01adjustment.

Delay in Starting at Pasadena

Because of the flood, there was over a two-hour delay before the Pasadena laptoppers could start the Thursday morning test session. To investigate whether this delay affected scores, we constructed a regression equation to predict an applicant's total score on essays 4–6 on the basis of that candidate's MBE, PT-A, essay 1–3 scores, standardized stall time, and whether the applicant did or did not take essays 4–6 at the Pasadena laptop center.

This analysis found that although the coefficient for answering essays 4–6 on a laptop at Pasadena (.0743) was positive (i.e., beneficial), it was not even close to being statistically significantly different than zero. In other words, the data indicate that the Pasadena laptoppers did about as well on essays 4–6 as would be expected given their scores on the rest of the exam (see Model 4 in Appendix B). There is certainly no evidence that the delay depressed their Thursday morning scores. Hence, there is no statistical basis for recommending that the scores at this center should be modified as a result of the delay.

Imputing PT-B Scores

The flood at the Pasadena test center led to canceling the Thursday afternoon PT-B session for the laptoppers at this site. We examined two methods, regression and pro-rata, for imputing the missing PT-B scores for the affected candidates. Some imputation is required because all applicants need a full complement of scores in order to compute their written and total scores; and thereby determine their pass/fail status.

Regression Imputation. The regression method used all of the 3745 applicants in our analysis sample that did not have their PT-B session cancelled. We used these applicants' data to construct a regression equation to predict their PT-B scores on the basis of their MBE scores and the sum of their scores on PT-A and essay questions 1–6 (after the stall time adjustment described above was implemented). This equation was as follows (the estimated PT-B score is expressed on the 100-point scale):

$$\text{Regression PT-B} = 23.31 + (0.0689)\,[(2 \times \text{PT-A}) + \text{Essays 1-6}] + (.0041)(\text{MBE})$$

We applied this equation to all of the 618 Pasadena laptoppers in the analysis sample. For example, a Pasadena laptopper with a PT-A score of 55 (which when multiplied by 2 is worth 110 points on the 1000-point scale) and an essay 1–6 score of 415 would have a total of 525 points on these two parts of the exam. If this applicant had an MBE scale score of 1347, then this applicant's imputed PT-B score would be 65 (see below):

Regression PT-B = 23.31 + (0.0689)[(2 x 55) + 415] + (.0041)(1347) = 65

Pro-rata Imputation. The pro-rata method assumes that a candidate's score on the 200 points allocated to PT-B is likely to be proportional to that applicant's scores on PT-A and essays 1–6. The pro-rata method uses a *scaling factor* to control for any differences in mean scores between the PT problems and essay questions. The formula below is used to compute this scaling factor*:*

Scaling Factor = [(2 x Mean on PT-A) + (Mean on essays 1–6)]/(Mean on PT-B)

This formula multiplies the PT-A scores by 2 so that the results are consistent with how an applicant's total written score (on the 1000-point scale) is computed. The means in this formula are computed after we apply the 0.01–point per second adjustment described above to the essay question 6 scores. In our sample of applicants who did not have their PT-B session cancelled, two times the mean on PT-A plus the mean on essays 1–6 was 497.78 points. Their mean on PT-B was 63.30. The scaling factor was therefore 497.78/63.30 = 7.86.

We also computed a scaling factor based on all of the 4057 candidates (including those taking the Attorneys' Exam) that had all eight written scores. In this group, two times the mean on PT-A plus the mean on essays 1–6 equaled 499.04 points. Their mean on PT-B was 63.49. These values also produced a scaling factor of 7.86. We therefore used the following pro-rata formula to impute a Pasadena laptopper's reader assigned PT-B score:

Pro-Rata PT-B score = [(2 x PT-A) + (essays 1–6)]/7.86

For example, a Pasadena laptopper with a reader assigned raw score of 55 on PT-A would have this score multiplied by 2. If this applicant earned 415 raw score points on essays 1–6, then this applicant would have an imputed PT-B score of 67 because [(2 x 55) + 415]/7.86 = 67. For score reporting purposes, all estimated values are rounded to the nearest whole number.

Comparison of Imputation Methods. We computed a total scale score for each of the 618 Pasadena laptoppers in our analysis sample using their regression imputed PT-B score and again using their pro-rata imputed PT-B score. Next, to simulate the pass/fail decision, we computed the percentage of these 618 applicants that had total scale scores of 1440 or higher under each imputation method after the first reading of their answers.

This analysis found that the two methods for imputing PT-B scores led to identical passing rates. These methods also agreed almost perfectly in which candidates they would pass and fail. The regression method "passed" one candidate that the pro-rata method "failed" and the pro-rata method "passed" one candidate that the regression method "failed." These results are consistent with those obtained in our modeling of these methods with the February 2003 data.

Conclusions and Recommendations

Our analysis of the stalling indicated that it did not result in any overall adverse effect on essay 4–6 scores. There was simply not much if any evidence of harm. If there was an adverse effect, it was limited to those relatively few candidates who averaged more than 5 minutes of stall time per hour of testing time (and just as many of them passed without an adjustment as those who had shorter standardized stall times). Hence, if an adjustment is to be made, we recommend using the one indicated by regression Model 3; i.e., adding 0.01 points to an applicant's score on essay question 6 for each second of average standardized stall time over 300 seconds that the candidate experienced. In addition, there was no statistical evidence that essay 4–6 scores were adversely affected by switching from using a laptop to handwriting on Thursday morning or by the delay in starting the Thursday morning session at the Pasadena laptop test center.

We recommend using the pro-rata method to impute PT-B scores for the Pasadena laptoppers who had this test session cancelled. The major advantage of this method over the regression approach is that it can be used with applicants who took the Attorneys' Exam; i.e., it does not require a separate formula for them.[2] In addition, the pro-rata method is consistent with the procedures the Committee of Bar Examiners used several years ago to estimate missing afternoon MBE scores when an earthquake required canceling that portion of the MBE. Given these considerations and the fact that the two imputation methods yield virtually identical results, we recommend that the Committee adopt the pro-rata method (with the 7.86 scaling factor). We also recommend that the imputation be based on the average of the first and second read scores on essays 1–6 and PT-A for those Pasadena laptoppers who went to reread.

Finally, in keeping with generally accepted measurement principles as described in the joint standards of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, we recommend that this imputation be made for just the candidates who were missing PT-B scores as a result of the flood. All other candidates should have their scores computed in the normal fashion.

[2] Attorneys who have practiced in another jurisdiction for at least four years can pass the California exam without having to take the MBE. Hence, their PT-B scores cannot be imputed with the same regression equation that was used for other applicants because that equation requires that the applicant have an MBE score.

Appendix A – Expert Panel

The expert panel provided advice on the analytic approach used and reviewed this report and its recommendations and conclusions prior to its release. Prof. Freedman also independently ran and corroborated all the regression equations reported in Appendix B.

Richard Berk is a professor in the Departments of Statistics and Sociology at UCLA. He also holds the position of Visiting Faculty Member at the Los Alamos National Laboratories. Professor Berk is an elected fellow of the American Association for the Advancement of Science and the American Statistical Association and has served on the Committee on Applied and Theoretical Statistics of the National Research Council, the National Center for Atmospheric Research Scientific Advisory Committee for the Climate Modeling Program, and the Social Science Research Council's Board of Directors. He has been awarded the Paul F. Lazarsfeld Award by the Methodology Section of the American Sociological Association. Professor Berk is a founding editor of the *Evaluation Review*, a position that he still holds. Professor Berk has published 13 books and over 150 peer reviewed articles and book chapters.

David A. Freedman received his B. Sc. degree from McGill and his Ph. D. from Princeton. He is professor of statistics at U. C. Berkeley, and a former chairman of the department. He has been Sloan Professor and Miller Professor, and is now a member of the American Academy of Arts and Sciences. He has written several books, including a widely used elementary text, as well as many papers in probability and statistics. He has worked on martingale inequalities, Markov processes, de Finetti's theorem, consistency of Bayes estimates, sampling, the bootstrap, census adjustment, procedures for testing and evaluating models, statistics and the law. In 2003, he received the John J. Carty Award for the Advancement of Science from the National Academy of Sciences. He has worked as a consultant for the Carnegie Commission, the City of San Francisco, and the Federal Reserve, as well as several Departments of the U. S. Government—Energy, Treasury, Justice, and Commerce. He has testified as an expert witness on statistics in a number of law cases, including Piva v. Xerox (employment discrimination), Garza v. County of Los Angeles (voting rights), and NewYork v. Department of Commerce (census adjustment).

Edward H. Haertel received his Ph.D. from the University of Chicago in 1980. After one year at the University of Illinois at Chicago Circle, he moved to Stanford University, where he is now a Professor in the School of Education. Haertel's research and teaching focus on theory, practice, and policy in educational testing and assessment, including test-based accountability and the use of test data for educational program evaluation. His recent work has focused on standard

setting and the validation of standards-based score reports and decision rules, and he is currently investigating the relation between accountability testing and "opportunity to learn." Haertel has served as president of the National Council on Measurement in Education (1998-99), as a member of the National Assessment Governing Board (1997-2003), and co-chairs advisory committees concerned with California's test-based school accountability system (1999-present). Haertel also served on the joint committee responsible for revising the Standards for Educational and Psychological Testing (1993-1999) and has also served on numerous state and national advisory committees related to educational testing, assessment, and evaluation. He has received the Cattell early career award as well as the Palmer O. Johnson award from the American Educational Research Association. Haertel is a Fellow of the American Psychological Association and a member of the National Academy of Education.

Robert Linn is Distinguished Professor of education in the research and evaluation methods program. Dr. Linn's research explores the uses and interpretations of educational assessments, with an emphasis on educational accountability systems. His work has investigated a variety of technical and policy issues in the uses of test data, including alternative designs for accountability systems and the impact of high-stakes testing on teaching and learning. His teaching interests are in related areas of educational measurement and statistical analysis. He has published more than 200 journal articles and book chapters dealing with a wide range of testing and assessment issues.

Dr. Linn received his AB from UCLA in 1961 and his MA and PhD in Psychology from the University of Illinois at Urbana-Champaign in 1964 and 1965, respectively. Dr. Linn is a member of the National Academy of Education and a lifetime National Associate of the National Academies. He has been an active member of the American Educational Research Association (AERA) for nearly 40 years and has served as vice president of the AERA Division of Measurement and Research Methodology and vice chair of the joint committee that developed the 1985 Standards for Educational and Psychological Testing. He is the current president of AERA. He is a past president of the National Council on Measurement in Education (NCME), past editor of the Journal of Educational Measurement, and editor of the third edition of Educational Measurement, a handbook sponsored by NCME and the American Council on Education. He was chair of the National Research Council's (NRC) Committee on Testing and Assessment and currently serves on the NRC's Board of the Center for Education.

Appendix B
Regression Coefficients (and p-values)

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	57.8149	58.3155	57.7992	58.3140
Essays 1–3 total score	0.3121 (.0001)	0.3116 (.0001)	0.3129 (.0001)	0.3116 (.0001)
PT-A	0.0269 (.3710)	0.0249 (.4067)	0.0266 (.3750)	0.0250 (.4054)
MBE	0.0458 (.0001)	0.0454 (.0001)	0.0456 (.0001)	0.0453 (.0001)
Used a laptop	-0.3616 (.4739)			
Standardized stall time (SST)		0.0023 (.2978)		0.0022 (.3539)
SST – 300 (if a positive value)			-0.0101 (.1095)	
Took exam at Pasadena				0.0743 (.9135)

Notes: All the models had adjusted R-squares of .33 and they were based on all of the applicants who had essay 1–6, PT-A, and MBE scores. Model 3 multiplies the difference between an applicant's standardized stall time and 300 seconds by –0.01 if that difference is greater than zero. The variable "took exam at Pasadena" in Model 4 was coded "1" if the applicant was a Pasadena laptopper. All other applicants were coded "0" for this variable.

Example: Suppose an applicant had an essay 1–3 score of 195, a PT-A score of 60, and an MBE score of 1400. According to Model 1, this applicant would have a predicted Essay 4–6 score of 183.97 if that applicant was a laptopper (see equation below) and a predicted score of 184.33 if the applicant was not a laptopper (because 0 times 0.3616 equals 0). The 0.36-point difference between these two estimates was not statistically significant.

$$y = 57.81 + (.3121 \times 195) + (.0269 \times 60) + (.0458 \times 1400) – (0.3616 \times 1) = 183.97$$

We also fitted a model that included a linear term for standardized stall time. The predicted essay 4–6 scores from this model increased with standardized stall time, which suggested that stall time was actually beneficial (see Models 2 and 4). When quadratic and cubic terms for standardized stall time were added to the model, predicted essay 4–6 scores increased and then decreased as standardized stall time increased. In short, there was no consistent evidence of adverse effects on the laptoppers' essay 4–6 scores as stall time increased.

Appendix C

Statistical Analyses of February 2003 Data

We conducted analyses with February 2003 exam data in preparation for our analyses of the February 2004 data. One of these analyses involved constructing a regression equation to predict essay 4–6 scores on the basis of the applicants' MBE, essay 1–3, and PT-A scores. This model also included a term for whether the candidate used a laptop or not. This analysis found that after controlling on the other variables in the model, using a laptop was not statistically significantly related to essay 4–6 scores (the coefficient for using a laptop was 0.37 and it had a p-value of .44). This finding of no significant effect is consistent with the one we obtained with the February 2004 data.

We also used the February 2003 data to construct a regression equation to predict the non-laptoppers' essay 4–6 scores on the basis of their MBE, PT-A, and essay 1–3 scores. We then applied this equation to the February 2003 laptoppers' MBE, PT-A, and essay 1–3 scores to estimate their essay 4–6 scores. This analysis found that 91 percent of the February 2003 laptoppers would have the same pass/fail status regardless of whether their predicted or actual essay 4–6 scores were used to compute their total scores.[3] Thus, the predicted scores from the regression analysis were found to be a good predictor of an applicant's pass/fail status.

The February 2003 data illustrated that it was not unusual to obtain a two or three point difference in mean question scores between essay test sessions. On the February 2003 exam, the Tuesday and Thursday morning means were 58.6 and 60.6, respectively. There was a 2.7-point difference in mean question scores between sessions on the February 2002 exam. Thus, the 3-point difference in mean essay question scores between sessions on the February 2004 exam was not an aberration.

Finally, we used the February 2003 data to compare the pro-rata and regression methods for imputing PT-B scores. This analysis found that these methods yielded virtually the same passing rates and resulted in the same pass/fail decisions for 97 percent of the 516 February 2003 Pasadena laptoppers. There was a 99.7 percent agreement rate in pass/fail decisions between these methods on the February 2004 exam.

[3] To facilitate comparisons, all the analyses in this report define "passing" as having a total scale score of 1440 or higher after the first reading of the applicants' answers.